D. Demmler, D. Krupka, H. Federrath. (Hrsg.): INFORMATIK 2022,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2022 651

Natural Language Processing (NLP) und der Datenschutz - Chancen und Risiken für den Schutz der Privatheit

Inna Vogel[1], Tahireh Setz[2], Jeong-Eun Choi[3], Prof. Dr. Martin Steinebach[4]

Abstract: Maschinelle Lernverfahren können sowohl Chancen als auch Risiken für die Privatheit von Daten bedeuten. Zum einen können durch Techniken des Natural Language Processings personenbezogene Daten anonymisiert werden und zum anderen können maschinelle Lernmodelle selbst hinsichtlich der Identifizierbarkeit der darin enthaltenen Daten zum Risiko für die Anonymität werden. In dieser Arbeit werden beide Aspekte, auch im Kontext von Angriffen auf die KI, diskutiert und Lösungsansätze besprochen. Sodann wird die datenschutzrechtliche Dimension von Angriffen auf die KI dargestellt und relevante Vorschriften des Entwurfs der Künstlichen-Intelligenz-Verordnung beleuchtet. Denn auch hier stehen sich der datenschutzrechtliche Grundsatz der Datenminimierung und das Interesse am Erhalt der Datenqualität gegenüber - ein scheinbar dilemmatisches Verhältnis.

Keywords: Textverarbeitung, Natural Language Processing (NLP), Datenschutz, Privatheit, Maschinelles Lernen.

1 Einleitung

Ob in der Textübersetzung, Bilderkennung oder bei der Prozessoptimierung in der Industrie: das maschinelle Lernen (ML) hat sich in vielen Bereichen bewährt. Der Siegeszug des ML kann mit der gestiegenen Rechenleistung sowie den frei verfügbaren Datenmengen begründet werden. Allerdings müssen die datenschutzrechtlichen Regelungen für die Verarbeitung der Daten beachtet werden. Die Daten für das Training von ML-Verfahren enthalten in der Regel personenbezogene Informationen (z.B. Namen, Zitate natürlicher Personen, Adressen), bei deren Verarbeitung die Anforderungen der Datenschutzgrundverordnung (DSGVO) einzuhalten sind. Werden Personenbezüge entfernt, ist der Verarbeiter – zumindest rechtlich – freier in den Nutzungsmöglichkeiten. Für die Entwicklung von ML-Verfahren gibt es aus technischer Sicht zwei Möglichkeiten: die Verschleierung oder Entfernung personenbezogener Daten. Diese technischen Vorgänge korrespondieren mit den rechtlichen Konzepten der Pseudonymisierung und Anonymisierung, wobei sie nicht gleichzusetzen sind (zu den Unterschieden s. Kap.2.1).

In dieser Arbeit werden verschiedene Techniken der maschinellen Textverarbeitung (engl. Natural Language Processing, kurz „NLP“) zum Schutz der Privatheit vorgestellt. Hierzu gehört die Entitätenerkennung, Koreferenzauflösung sowie die Schreibstilverschleierung. Allerdings können auch ML-Verfahren selbst hinsichtlich der darin enthaltenen Daten

[134] Fraunhofer SIT, Rheinstraße 75, Darmstadt, 64295, [Vorname].[Nachname]@sit.fraunhofer.de
[2] Universität Kassel, Wirtschaftswissenschaften, Henschelstraße 4, Kassel, 34127, t.setz@uni-kassel.de

doi:10.18420/inf2022_52

zum Risiko für die Anonymität werden. Welche Schutzmechanismen es gegen Angriffe auf ML-Modelle gibt, wird in Kapitel 2.3 diskutiert. Kapitel 3 widmet sich datenschutzrechtlichen Aspekten. Hier wird ein rechtlicher Bezug zu den vorgestellten NLP-Verfahren sowie den Angriffen auf ML-Systeme und Schutzmechanismen hiergegen hergestellt. Der Verlust der Nutzbarkeit der Daten für Forschung und Praxis steht den Anforderungen des Datenschutzes gegenüber. Dieses scheinbar dilemmatische Verhältnis wird zum Schluss der Arbeit erläutert und Lösungsansätze dafür vorgestellt.

2 NLP und der Datenschutz

Maschinelle Lernverfahren bieten sowohl Chancen als auch Risiken für die Privatsphäre. Zum einen können ML-Verfahren zur Anonymisierung personenbezogener Daten genutzt werden und zum anderen können diese selbst hinsichtlich der Anonymität der darin enthaltenen Daten zum Risiko für die Anonymität werden. Beide Aspekte und wie der Datenschutz schon während der Datensammlung berücksichtigt werden kann, werden nachfolgend vorgestellt.

2.1 Anonymisierung/Pseudonymisierung während der Datensammlung

Um Informationen aus verschiedenen Interneträumen zu erfassen, werden sogenannte „Crawler“ eingesetzt. Hierbei handelt es sich um Programme, die sich automatisiert über Verlinkungen (URLs) auf Webseiten durch das Internet bewegen, Inhalte erfassen und herunterladen. Die Inhalte können nicht nur HTML-Webseiten, sondern auch andere Datenformate wie z.B. Text-, Bild-, Audio- und Videodateien sein. Als „Scraper“ wird diejenige Teilkomponente eines Crawlers bezeichnet, welche die Analyse der erfassten Inhalte übernimmt. Einerseits müssen Verlinkungen für die nächsten Webseiten identifiziert und erfasst werden, andererseits muss eine Selektion der Inhalte auf Webseiten ermöglicht werden, um Daten zu erfassen, welche für das jeweilige Projekt relevant sind. Da Webseitenstrukturen individuell sind, wird für jede Webseite ein eigener Scraper benötigt, um die relevanten Inhalte adressieren zu können. Das bedeutet aber auch, dass bereits beim Crawlen die Verarbeitung personenbezogener Daten beschränkt und das Prinzip der Datenminimierung vorgenommen werden kann (Art. 5 Abs. 1 lit. c DSGVO). Zu den technisch-organisatorischen Maßnahmen der Datenminimierung kann die Pseudonymisierung gehören (Art. 25 Abs. 1 DSGVO). Demgegenüber ist auch die Anonymisierung stets zu prüfen. Anonymisierung meint die Veränderung personenbezogener Daten, sodass die Einzelangaben über persönliche oder sachliche Verhältnisse nicht mehr oder nur mit einem unverhältnismäßigen Aufwand an Zeit, Kosten und Arbeitskraft einer bestimmten oder bestimmbaren natürlichen Person zugeordnet werden können, siehe [RG21, S. 487]. Die in Art. 4 Nr. 5 DSGVO legal definierte Pseudonymisierung bewirkt die Funktionstrennung von Zuordnungs-information und Daten [Sc20, S. 284, 285].

Werden beispielsweise soziale Netzwerke wie Facebook gecrawlt, können personenbezogene Daten wie Profilname und -bild, Geburtsdatum, Wohnort, Namen der Follower und das Profilbild vom Crawler ignoriert werden, um den Datenschutz zu gewährleisten. Auf diese Weise kann die Weiterverarbeitung der Daten für Forschungszwecke ohne Personenbezug erfolgen (Art. 89 Abs. 1 Satz 4 DSGVO). Werden allerdings bestimmte Nutzerinformationen benötigt, um beispielsweise Korrelationen zwischen Usern zu erkennen, können personenbezogene Daten auch während des Crawlens pseudonymisiert werden und zusätzlich verschlüsselt gespeichert werden [St20]. Dennoch muss erwähnt werden, dass die Anonymisierung oder Pseudonymisierung von öffentlich zugänglichen Daten zwar eine risikominimierende Wirkung hat, die Zuordnung der betroffenen Person jedoch nicht ausgeschlossen werden kann. Um die Rückführung der Autorschaft zu minimieren, können stilverschleiernde Methoden in Betracht gezogen werden (siehe Kapitel 2.2).

2.2 Anonymisierung/Pseudonymisierung bei der Verarbeitung der erfassten Daten

Anders als bei strukturierten Daten (z.B. HTML-Code), wenn jedem Wert eine Kategorie zugeordnet werden kann, besitzen unstrukturierte Daten kein festes Schema oder kategorische Werte. Um schützenswerte Daten in unstrukturierten Freitextdokumenten (wie z.B. in Briefen, E-Mails, OP-Berichten oder Social Media-Kommentaren) zu anonymisieren/pseudonymisieren, können verschiedene Techniken des NLP angewandt werden. Um Entitäten wie Personen, Organisationen, Orte und numerische Daten (wie z.B. Geburtsdatum, Telefonnummer, Körpermaßeinheiten etc.) automatisiert zu identifizieren, kann die Eigennamenerkennung (engl. Named-Entity-Recognition, kurz „NER“) eingesetzt werden. Die besten Ergebnisse erzielt momentan „Flair“[5], ein Modell von Hugging Face, welches auf einem vortrainierten Transformer basiert, einer 2019 vorgestellten Deep-Learning-Architektur [SA20, Ak19]. Das Verfahren wurde auf den CoNLL-2003 Datensätzen in verschiedenen Sprachen trainiert [SA20]. Bei einer Vier-Klassen-Erkennung[6] werden F1-Werte für die deutsche Sprache von 92,31 erreicht, für die englische Sprache 94,36 (siehe Tab. 1).

	CoNLL-2003 (Deutsch) [SA20]	CoNLL-2003 (Eng.) [SA20]	Twitter (Deutsch) [St19]
FLAIR	92,31	94,36	82,01

Tab. 1: F1-Score von FLAIR [SA20], eigene Evaluierung auf Twitter-Posts [St19]

Die Genauigkeit bei der Erkennung von Personennamen sinkt, wenn das Verfahren auf Twittertexten angewandt wird. Eine manuelle Evaluierung auf 400 Twitter-Posts hat einen

[5] Flair-Modell von Hugging Face: https://huggingface.co/flair/ner-german-large
[6] NER Vier-Klassen: Person, Organisation, Ort und „Miscellaneous“ (alle weiteren Eigennamen)

F1-Wert von 82,01 ergeben (siehe Tab. 1) [St19]. Das liegt zum einen daran, dass bei Twitter weniger auf Rechtschreibung und Grammatik geachtet wird, aber auch daran, dass oft keine Standardsprache verwendet wird. Weiterhin gelten plattformspezifische Regeln wie die Verwendung von Hashtags oder @-Mentions, was eine Entitätenerkennung zusätzlich erschwert.

Einen höheren Datenschutz bietet die Koreferenzauflösung (engl. Coreference Resolution), welche zur Aufgabe hat, Relationen zwischen Textelementen zu finden, die auf dieselbe Entität verweisen. Beispielsweise verweist der Ausdruck „*die ehemalige Bundeskanzlerin*“ auf die Person „*Angela Merkel*“. Aber auch Pronomen (z.B. „*sie*“, „*seine*“) und andere Ausdrücke können referierend sein (z.B. verweist „*Onlineversand-händler*“ auf „*Amazon"*) [LHZ18]. Coreference Resolution kann eingesetzt werden, um einen höheren Datenschutz zu gewährleisten. So werden nicht nur Entitäten erkannt, sondern auch alle Begriffe, die sich auf die Entität beziehen, z.B. Berufsbezeichnungen und nähere Personenangaben wie „*die Krebspatientin*“, „*der Schlosser*“ oder „*der in Marburg lebende Jurastudent*“. Garat und Wonsever [GW22] haben die Koreferenzauflösung z.B. dazu eingesetzt, um spanische Gerichtsakten zu anonymisieren. Ziel war es nicht nur Personennamen, sondern auch alle Koreferenzen aufzulösen, um alle Erwähnungen derselben Person einheitlich zu pseudonymisieren. Die Schwierigkeit hierbei war ein Verfahren zu entwickeln, welches in der Lage ist unterschiedliche Schreibweisen von Personennamen zu erkennen. Hierfür wurde ein NER-Verfahren (F1-Score 90,21) zusammen mit einem Clusteringalgorithmus eingesetzt, um Koreferenzen aufzulösen (Accuracy 81). Allerdings merken die Autoren an, dass auf der Validierungsmenge die Genauigkeit bei der Erkennung und Verknüpfung der Namen mit ihren Koreferenzen immer noch unter 50% liegt. Generell lässt sich festhalten, dass Coreference Resolution-Verfahren niedrigere Genauigkeitswerte erzielen als NER-Verfahren. Das deutsche „E2E“[7]-Verfahren wurde auf verschiedenen Datensätzen (Literatur- und Nachrichtentexten) getestet und erzielt Genauigkeitswerte zwischen 64,72 (DROC-Datensatz) und 80,2 (CoNLL-Datensatz von 2012). Das CorefQA-System von Wu et al. [We20] ist derzeit das beste System für Englisch und erzielt auf dem CoNLL-2012 Datensatz einen F1-Score von 83,1 (siehe Tab. 2).

	TüBa (Deutsch)	SemEval (Deutsch)	DROC (Deutsch)	CoNLL-2012 (Eng.)
E2E (Deutsch)	78,79	74,46	64,72	80,2
CorefQA (Eng.)	-	-	-	83,1

Tab. 2: F1-Score für verschiedene Coreference Resolution-Verfahren [SHB21, We20]

Selbst wenn personenbezogene Daten wie Name und E-Mail-Adresse im Text anonymisiert werden, kann der Autor anhand seines Schreibstils mit sogenannten

[7] „Neural End-to-end Coreference Resolution for German in Different Domains“

Autorschafts-Attributionssystemen (engl. Authorship Attribution) identifiziert werden. Um die Anonymität des Verfassers dennoch zu wahren, können Verfahren eingesetzt werden, die den Schreibstil im Text verfälschen (engl. Authorship Obfuscation). Ziel der Autorenverschleierung ist es, einen gegebenen Text automatisch so zu paraphrasieren, dass dieser weder von Menschen noch von modernen Verfahren zur Überprüfung der Autorschaft (engl. Authorship Verification) dem ursprünglichen Autor zugeordnet werden kann [Be20]. Es zeigt sich allerdings, dass viele Verfahren den Text so verändern, dass entweder der Kontext (d.h. die Semantik) zu stark verändert wird oder der Text gar unlesbar für den Menschen wird [HPS17, MSS20]. Mahmood et al. [MSS20] haben beispielsweise ein Verfahren vorgestellt, welches mit einem durchschnittlichen F1-Wert von 87 erkennen kann, ob ein Text verändert wurde oder nicht. Sie geben an, dass bestehende Methoden zur Verschleierung der Autorschaft selbst stilistische Signaturen hinterlassen, die mit Sprachmodellen identifiziert werden können. Ihre Ergebnisse verdeutlichen die Notwendigkeit, Methoden zur Verschleierung der Autorschaft weiterzuentwickeln, welche die Identität eines Autors, der Anonymität anstrebt, besser schützen.

2.3 Risiken für die Privatheit von maschinellen Lern- und Sprachmodellen

Trainierte maschinelle Lernmodelle sowie das Training dieser Modelle auf Daten mit Personenbezug bergen immer Risiken für die Privatheit. Lange wurde die Ansicht vertreten, dass ML-Verfahren die Anonymisierung von Trainingsdaten garantieren, da die Datensätze, die für das Training der ML-Algorithmen verwendet werden, in der Regel disjunkt mit den Datensätzen sind, welche in der Anwendung eingesetzt werden. Nur vom Modell Rückschlüsse auf die Trainingsdaten zu ziehen, sollte somit nicht möglich sein, was einer Anonymisierung der Trainingsdaten gleichkäme [WBH19, BG21]. Attacken auf ML-Modelle wie die „Membership Inference“ verwenden Datenpunkte, um zu rekonstruieren, ob diese für das Training des Modells verwendet wurden. Bei „Model-Inversion-Attacken“ können Angreifer zum einen für einen bestimmten Datenpunkt personenbezogene Attribute rekonstruieren, zum anderen sogar ganze Teile der Trainingsdaten wiederherstellen. Solche Angriffe stellen eine Verletzung der Privatheit dar, vor allem, wenn es um besondere Kategorien der personenbezogenen Daten nach Art. 9 Abs. 1 DSGVO geht und eine Anonymisierung ausdrücklich erforderlich ist. Solche Informationen im Trainingsdatensatz können beispielsweise mittels „Distillation“ geschützt werden. Hierbei wird ein Lernmodell verwendet, um einen weiteren Datensatz zu klassifizieren. Dieser neu erstellte Datensatz wird im Anschluss dazu verwendet, um ein neues separates Modell zu trainieren. Hierfür wird das Wissen des ursprünglichen Modells übertragen [Pa18]. Es ist allerdings zu beachten, dass alle Maßnahmen zum Schutz der Privatheit mit einer Reduktion der Klassifikationsgenauigkeit einhergehen [BG21].

Sprachmodelle (engl. Language Models) wie GPT-3[8] von OpenAI oder PaLM[9] von Google sind darauf trainiert, die Wahrscheinlichkeit aufeinanderfolgender Wörter im Satz zu berechnen und auf diese Weise unter minimalen Vorgaben Texte zu ergänzen und selbst zu verfassen. Die Grundlage der Sprachmodelle sind Hunderte von Gigabyte an Textdaten aus dem Web wie etwa E-Books, Wikipedia, Social-Media-Plattformen oder Reddit. Zwar haben die Sprachmodelle zu bahnbrechenden Verbesserungen in vielen NLP-Anwendungen – wie etwa der Textgenerierung oder -übersetzung – geführt, können allerdings sensible Unternehmens- und Kundendaten (wie Namen, Telefonnummern, Adressen usw.) preisgeben. Carlini et al. [Ca21] führten einen Proof-of-Concept-Angriff zur Extraktion von Trainingsdaten auf GPT-2 durch und zeigten, dass es möglich ist, die genannten privaten Informationen aus dem Modell zu extrahieren. Um dieses Problem zu minimieren, wäre ein Training der Sprachmodelle auf anonymisierten oder unproblematischen Daten denkbar, was bei der benötigten Datenmenge jedoch schwierig umzusetzen ist. Eine weitere Möglichkeit wäre der Einsatz von „Differential Privacy". Hierbei handelt es sich um ein Konzept, Datenmengen zu analysieren, ohne die Privatsphäre eines einzelnen Benutzers zu verletzen, indem die Originaldaten mithilfe von Hash-Verfahren (hashing), Anreicherung mit Zufallswerten (noise injection) und der Verwendung von Teilmengen (subsampling) so verändert werden, dass sie keine Rückschlüsse auf personenbezogene Daten mehr zulassen [Dw06]. Das Framework DPText [Be19] soll beispielsweise das Trainieren von Textrepräsentation ermöglichen, welche keine Rückschlüsse auf private Informationen ermöglichen. Der Algorithmus ADePT [KGD21] paraphrasiert den gegebenen Text und garantiert auf diese Weise Different Privacy im neu generierten Datensatz. Habernal zeigte jedoch, dass weder DPText noch ADePT Differential Privacy garantieren können [Ha21, Ha22]. Weiter merkt er an, dass es bei einem strengeren Datenschutz durchaus zu einem bedeutenden Performanzverlust kommt und die Wahl differenzierter Datenschutzmaßnahmen nach Bedarf an Datenschutz und Performanz beurteilt werden sollte [SIH21]. Auch werden durch die auf dem Konzept der Differential Privacy beruhenden Techniken die Originaldaten als solche nicht verändert und gelten rechtlich weiterhin als personenbezogene Daten [SB21][10]. Zudem werden durch das Verzerren der Daten Veränderungen verursacht, die zu einer Minderung der Datenqualität führen können. Sinnvoll könnte eine kombinierte Differential-Privacy-Garantielösung sein, wonach eine datenschutzrechtliche Garantie eine Voraussetzung ist, das Modell ohne Rauschen zu nutzen. Ohne Garantie ist nur eine durch Differential Privacy verrauschte Version erhältlich. Eine datenschutzrechtlich vorteilhafte Variante der Differential Privacy stellt das Jittering dar. Hierdurch werden Daten lediglich punktuell verzerrt, was – im Gegensatz zur Generalisierung/Randomisierung – nur zu einer minimalen Verzerrung des Gesamtbildes führt. Dafür werden ebenjene Einzeldaten ausgewählt, deren Änderungen relevant für die Privatsphäre sind. Durch eine präzise und geschickte Auswahl lässt sich

[8] GPT-3: https://openai.com/blog/openai-api/

[9] Pathways Language Model (PaLM): https://ai.googleblog.com/2022/04/pathways-language-model-palm-scaling-to.html

[10] Bretthauer, in: Spiecker gen. Döhmann/Bretthauer, Dokumentation zum Datenschutz, G. G 2 G 2.1 G 2.1.10 Stellungnahme 05/2014 zu Anonymisierungstechniken (WP 216) 3 3.1. 3.1.3.

sowohl eine repräsentative Genauigkeit der Ergebnisse als auch der Schutz der Privatsphäre erhalten, siehe S. 329, 382 in [Pa20].

3 Datenschutzrecht

Die maschinelle Textverarbeitung personenbezogener Daten muss sich im Rahmen der DSGVO bewegen. Der in Art. 5 Abs. 1 lit. c DSGVO verankerte Grundsatz der Datenminimierung kann u.a. durch die technisch-organisatorische Maßnahme der Pseudonymisierung verwirklicht werden. Jedoch kann sowohl durch Anonymisierung als auch Pseudonymisierung ein erheblicher Verlust der Datenqualität und damit Nutzbarkeit in Forschung und Praxis entstehen.

Gegenüber der Pseudonymisierung ist die Anonymisierung, die nach EG Nr. 26 S. 5 DSGVO die Anwendbarkeit der DSGVO entfallen lässt, grundsätzlich vorrangig anzuwenden [SHS19, Art. 89 DSGVO, Rn 51 f.]. Die Pseudonymisierung findet an verschiedenen Stellen der DSGVO Ausdruck, z.B. Art. 25 Abs. 1 und Art. 32 Abs.1 a). Für die privilegierte Datenverarbeitung zu wissenschaftlichen Zwecken gilt nach Art. 89 Abs. 1 DSGVO, dass technisch-organisatorische Maßnahmen wie die Pseudonymisierung zu treffen sind. Für besondere Kategorien personenbezogener Daten i.S.d. Art. 9 Abs. 1 DSGVO sieht das deutsche Recht in § 27 Abs. 3 BDSG Anonymisierung unter bestimmten Voraussetzungen ausdrücklich vor.

Pseudonyme Daten können – je nachdem, ob eine Person über die Zuordnungsinformation verfügt oder nicht - anonym oder personenbezogen sein [Ro18, S. 243, 246]. Gerade die anonymisierende Pseudonymisierung, bei der sämtliche Bezugspunkte zur Person aus den Daten entfernt werden, kann die Datenqualität jedoch erheblich beeinträchtigen. Zum Ausgleich dieses Spannungsverhältnisses kann hingegen die risikomindernde Pseudonymisierung nach Art. 32 Abs. 1 lit. a DSGVO von Vorteil sein [Vgl. Ro18, S. 245 f.]. In der nach Art. 32 Abs. 1 lit. a, Abs. 2 DSGVO zu erfolgenden Abwägungsentscheidung steht das Interesse des Datenverarbeiters am Erhalt qualitativ ergiebiger Daten den Grundrechten der natürlichen Personen, die von der Verarbeitung betroffen sein können, gegenüber. Denn im Rahmen der Beurteilung der Risiken und Folgen sind nach Art. 32 Abs. 2 DSGVO auch die mit der Verarbeitung selbst verbundenen Risiken des Verlusts einzubeziehen [Sy18, Art. 32 Rn. 10]. Gem. Art. 32 Abs. 1 S. 1 DSGVO ist der Zweck der Verarbeitung in der Risikoabwägung zu beachten. Werden Daten zu Forschungszwecken verarbeitet, kann der Schutz der Datenqualität stärker gewichtet werden als bei rein wirtschaftlichen Zwecken. Zwar steht dem Verantwortlichen und dem Auftragsverarbeiter sein Ermessen bei der Auswahl der Maßnahmen zu (Ebd.). Zu berücksichtigen ist aber die gesetzlich vorgegebene Zielrichtung der Abwägung: Die Risikobeurteilung nach Art. 32 DSGVO muss zum Zweck der Sicherstellung des angemessenen Schutzniveaus aus der Sicht der betroffenen Personen und nicht des Datenverarbeiters erfolgen [Sy18, Art. 32 Rn. 10; SHS19, Art. 32 DSGVO, Rn. 28].

3.1 Angriffe auf maschinelle Lernverfahren

Besonders zeigt sich das Problem des Verlusts der Datenqualität durch Pseudonymisierung bei maschinellen Lernmodellen. Einerseits hängt der wissenschaftliche bzw. wirtschaftliche Nutzen der Sprachmodelle [AK20, S. 24, 25] von der Qualität und Quantität der Trainingsdaten ab [KB20, S.462]. Andererseits fordert der Grundsatz der Datensparsamkeit nach Art. 5 Abs. 1 lit. c DSGVO ein Lernen mit so wenigen personenbezogenen Daten wie möglich [NK20, Rn 46], sodass eine ursprüngliche Anonymisierung oder ein Training mit wenigen Daten vorzuziehen wäre. Das besondere Risiko der Re-Identifikation besteht dadurch, dass durch Angriffe auf Sprachmodelle – entgegen bisheriger Annahmen - Rückschlüsse auf personenbezogene Trainingsdaten möglich sind (s.o.). Bei der Bestimmung der Identifizierbarkeit personenbezogener Daten sind gem. EG 26 zur DSGVO alle Mittel zu berücksichtigen, die von dem Verantwortlichen oder einer anderen Person nach allgemeinem Ermessen wahrscheinlich genutzt werden. Dabei sind auch Faktoren wie die Kosten der Identifizierung und der dafür erforderliche Zeitaufwand und die zum Zeitpunkt der Verarbeitung verfügbare Technologie und technologische Entwicklungen zu berücksichtigen.

Die technologische Entwicklung ermöglicht Angriffe gegen Sprachmodelle, die personenbezogene Trainingsdaten rekonstruieren. Ein besonderes Risiko der Identifizierung geht außerdem von Model-Inversion- und Inference-Attacken aus (siehe Kapitel 2.3), die deutlich machen, dass eine automatische Anonymisierung der Trainingsdaten nicht garantiert werden kann. Ob noch von einer faktischen Anonymität, die die Anwendbarkeit der DSGVO ausschließt, ausgegangen werden kann, ist indes fraglich [KB20, S.466 m.w.N]. Insbesondere ist zu beachten, dass Angriffe einen erheblichen Aufwand bedeuten. Zweifelhaft ist also, ob wegen des Aufwands von einer hinreichenden Wahrscheinlichkeit ausgegangen werden kann, dass Angriffe tatsächlich durchgeführt werden (Ebd., S. 467). Auch ist das Risiko, dass Rückschlüsse auf natürliche Personen gezogen werden, auf Grund der für Angriffe vorausgesetzten gewissen Kenntnis der Modelle als gering einzuschätzen, [KB20, S. 467] sodass jedenfalls eine faktische Anonymisierung angenommen werden kann, sofern im Einzelfall keine anderen Gründe entgegengehalten werden können.

Risiken durch wirtschaftlich oder anders motivierte Angriffe sind auch im Rahmen der Abwägung nach Art. 32 Abs. 1 lit. a, Abs. 2 DSGVO (s.o.) zu berücksichtigen. Abs. 2 nennt explizit auch Risiken durch unrechtmäßige Ereignisse, worunter auch Angriffe zählen [SHS19, Art. 32 DSGVO Rn. 59 m.w.N].

Schließlich wird vorgeschlagen, bestimmte Ausprägungen der willentlichen De-anonymisierung zu sanktionieren, z.B. durch Einführung eines Straftatbestandes auf nationaler Ebene [RG21, S. 487, 490].

3.2 KI-VO-E

Eine Lösung für das Problem des Verlusts der Datennutzbarkeit wird in Art. 10 Abs. 1 des EU-Kommissionsentwurfs für eine Verordnung über die Künstliche Intelligenz (KI-VO-E) vorgeschlagen. Dieser sieht vor, dass Hochrisiko-KI-Systeme i.S.d. Art. 6 KI-VO-E mit solchen Daten entwickelt werden müssen, die den in den Qualitätskriterien aus Abs. 2 - 5 entsprechen. Diese richten sich auch auf die Vollständigkeit der Daten. Abs. 2 zählt exemplarisch Schritte der Entwicklung und des Betriebs von Hochriko-KI-Systemen auf, bei denen geeignete Governance- und Datenverwaltungsverfahren gelten sollen. Dies betrifft nach Abs. 2 g) auch die Ermittlung von Datenlücken oder Mängeln und Methoden zur Behebung. Nach Abs. 3 S. 1 müssen Trainings-, Validierungs- und Testdatensätze relevant, repräsentativ, fehlerfrei und vollständig sein. Durch diese Governance-maßnahmen soll die Gefahr der Diskriminierung, die sich aus Verzerrungen in KI-Systemen ergeben könnte, gebannt werden (Vgl. EG 44).

Dass der Verordnungsgeber sich über die Risiken der Re-Identifizierung von Personen durch Rückgriff auf die Trainingsdaten bewusst ist, wird in EG 51 zur KI-VO-E deutlich. Hier wird von der Widerstandsfähigkeit von KI-Systemen gegenüber Versuchen böswilliger Dritter gesprochen, die die Schwachstellen der Systeme ausnutzen, indem sie z.B. auf Trainingsdatensätze zurückgreifen oder trainierte Modelle nutzen. Hiergegen sollten die Anbieter von Hochrisiko-KI-Systemen geeignete Maßnahmen ergreifen (z.B. IKT-Infrastruktur). Trotz des legitimen Zwecks sicherzustellen, dass Hochrisiko-KI-Systeme bestimmungsgemäß und sicher funktionieren und nicht zur Ursache für Diskriminierung werden (vgl. EG 44), würde der Erlass der vorgeschlagenen Vorschrift das Spannungsfeld zwischen Informationsgehalt und Datenschutz noch verstärken und zu einem dilemmatischen Verhältnis zwischen Erhalt der Datenqualität und dem Schutz personenbezogener Daten führen. Zum einen ist unklar, was unter „fehlerfrei“ zu verstehen ist [BM21, S. 276, 280; [RW21, S. 844, 847]. Die Gefahr besteht, dass ein während des Trainings hinzugefügtes Rauschen durch die zum Schutz personenbezogener Daten eingesetzte Differential Privacy (s. Kapitel 2.3) als Fehler gewertet werden könnte [Eb21, S. 528, 533], obwohl diese Maßnahme zum Schutz personenbezogener Daten sinnvoll wäre. Das Problem zeigt sich weiterhin in Abs. 5, wonach die Verarbeitung besonderer Kategorien personenbezogener Daten nach Art. 9 Abs. 1 DSGVO erlaubt ist, soweit es für die Beobachtung, Erkennung oder Korrektur von Verzerrungen in Hochrisiko-KI-Systemen unbedingt erforderlich ist (1.HS). Im 2. HS wird gefordert, dass angemessene Vorkehrungen für den Schutz der Grundrechte und Grundfreiheiten natürlicher Personen getroffen werden müssen, wozu auch die Pseudonymisierung gehört, wenn der verfolgte Zweck durch eine Anonymisierung erheblich beeinträchtigt würde. Offen bleibt, wie dieses Ziel zugleich mit der in Art. 10 Abs. 1 geforderten Vollständigkeit und Fehlerlosigkeit der Daten gewährleistet werden soll.

Bei Verstößen gegen Art. 10 KI-VO-E könnten nach Art. 71 III KI-VO-E Bußgelder bis zu 30.000.000 Euro oder 6% des weltweiten Jahresumsatzes verhangen werden. Bußgelder im Falle der willentlichen De-anonymisierung durch Angriffe gegen Modelle sind jedoch nicht vorgesehen.

4 Diskussion

Wir haben unterschiedliche Methoden der Datenanonymisierung mittels NLP-Methoden sowie während des Datensammelns (Crawling) vorgestellt. Wird die Koreferenzauflösung für die Anonymiserung von Daten verwendet, bei der sämtliche Bezugspunkte zur Person aus den Daten entfernt werden, kann die Datenqualität erheblich beeinträchtigt werden. NER könnte eine Lösung sein, welche mit weniger Informationsverlust einhergeht. Wenn auch der Textinhalt besonders schützenswert ist, können Techniken der Stilverschleierung eine Option sein. Dennoch bergen alle maschinellen Lernverfahren Gefahren, die beim Einsatz berücksichtigt werden sollten. Es existieren bislang keine ML-Verfahren, die eine hundertprozentige Zuverlässigkeit garantieren.

Zudem können ML-Verfahren selbst hinsichtlich der Anonymität der darin enthaltenen Daten zum Risiko werden, beispielsweise durch Membership-Inference oder Model-Inversion-Attacken. Um sich vor solchen Angriffen zu schützen, haben wir einige Ansätze wie Distillation, Differential Privacy und Jittering vorgestellt. Allerdings gehen diese Verfahren immer mit Performanz- und Datenqualitätsverlust einher. Und ob von einer faktischen Anonymität durch den Einsatz solcher Verfahren, die die Anwendbarkeit der DSGVO ausschließt, ausgegangen werden kann, ist indes fraglich. Andererseits ist das Risiko des Zugriffs auf personenbezogene Daten durch Angriffe auf ML-Systeme nach jetzigem Kenntnisstand als eher gering einzustufen. Denn solche Angriffe erfordern ein erhebliches technisches Wissen sowie eine begründete Motivation des Angriffs.

Der EU-Kommissionsentwurfs KI-VO-E sieht diesbezüglich vor, dass für Hochrisiko-KI-Systeme Datensätze entwickelt werden müssen, welche die folgenden Qualitätskriterien erfüllen: Vollständigkeit, Relevanz, Repräsentativität. Datenlücken und -mängel müssen geschlossen werden, wobei gleichzeitig von „*Widerstandsfähigkeit von KI-Systemen gegenüber Versuchen böswilliger Dritter*" gesprochen wird, was das Spannungsfeld zwischen Informationsgehalt und Datenschutz verstärkt. Zwar wird im KI-VO-E der Schutz besonderer Kategorien personenbezogener Daten und Systeme gefordert, was durch Einsatz von Differential Privacy ermöglicht werden könnte. Der Verordnungsentwurf lässt jedoch offen, wie zugleich eine Fehlerlosigkeit der Daten gewährleistet werden kann. Eine logische Schlussfolgerung wären NLP-Verfahren, welche nicht nur eine zuverlässige Datenanonymisierung bzw. -pseudonymisierung garantieren, sondern ebenfalls die Aufrechterhaltung der Datenqualität und Semantik. Auch der Schutz der Daten durch Jittering scheint eine minimalinvasive Methode zu sein, die sowohl eine hinreichende Datenqualität als auch Datenschutz gewährleistet [Pa20, S.376, 382].

5 Fazit

Die Entwicklung des maschinellen Lernens zeigt immer neue Risiken der Re-Identifizierung personenbezogener Daten und damit technische Grenzen der

Anonymisierung bzw. Pseudonymisierung auf. Gleichzeitig bringt der Einsatz stark anonymisierender bzw. pseudonymisierender Methoden die Gefahr des Verlusts der Nutzbarkeit der Daten für Forschung und Praxis mit sich. Hierdurch besteht ein Spannungsfeld zwischen dem datenschutzrechtlichen Prinzip der Datenminimierung, das den durch die Datenverarbeitung Betroffenen schützen soll, und dem Interesse des Verarbeiters an der Minimierung des Qualitätsverlusts. Dabei handelt es sich indes nur um ein scheinbar dilemmatisches Verhältnis: Gerade um die Sicherheit des Betroffenen zu gewährleisten (auch was den Schutz vor Diskriminierung durch Verzerrung von KI-Systeme angeht, vgl. EG 44 KI-VO-E), müssen der wissenschaftlichen Erforschung und anwendungsorientierten Entwicklung maschineller Lernmodelle hinreichend qualitative Daten zur Verfügung stehen. Dies wurde zwar auch in der KI-VO-E erkannt, dennoch müssen die Vorgaben des Art. 10 Abs. 1 KI-VO-E hinsichtlich der Datenqualität weniger streng sein, um einen ausreichenden Datenschutz zu gewährleisten. Dies gilt gerade auch dafür, um bessere datenschutzfreundliche technische Maßnahmen entwickeln zu können und andere grundrechtlich gebotenen Ziele (z.B. Vermeidung von Diskriminierung) zu erreichen.

Danksagung

Diese Forschungsarbeit wurde vom Bundesministerium für Bildung und Forschung (BMBF) im Rahmen des Projekts DYNAMO unterstützt.

Literatur

[Ak19] Akbik, A.; Bergmann, T.; Blythe, D.; Rasul, K.; Schweter, S.; Vollgraf, R.: FLAIR: An easy-to-use framework for state-of-the-art NLP. In: NAACL 2019, 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations). S. 54-59, 2019.

[AK20] Apel, S.; Kaulartz, M.: Rechtlicher Schutz von Machine Learning-Modellen. RDi 1/1, S. 24-34, 2020.

[Be19] Beigi, G.; Shu, K.; Guo, R.; Wang, S.; Liu, H.: Privacy Preserving Text Representation Learning. In: Proceedings of the 30th ACM Conference on Hypertext and Social Media. HT '19, Association for Computing Machinery, Hof, Germany, S. 275-276, 2019, url: https://doi.org/10.1145/3342220.3344925.

[Be20] Bevendorff, J.; Wenzel, T.; Potthast, M.; Hagen, M.; Stein, B.: On divergence-based author obfuscation: An attack on the state of the art in statistical authorship verification, it - Information Technology 62/2, S. 448-452, Juli 2021.

[BG21] Battis, V.; Graner, L.: Risiken für die Privatheit aufgrund von Maschinellem Lernen. In (Reussner, R. H.; Koziolek, A.; Heinrich, R., Hrsg.): INFORMATIK 2020. Gesellschaft für Informatik, Bonn, S. 841-855, 2021.

[BM21] Bomhard, D.; Merkle, M.: Europäische KI-Verordnung. Der aktuelle Kommissionsentwurf und praktische Auswirkungen, RDi 2/6, S. 276-283, 2021.

[Bo21] Boenisch, F.: Privatsphäre und Maschinelles Lernen. Über Gefahren und Schutzmaßnahmen, Datenschutz und Datensicherheit - DuD 45/7, S. 448-452, Juli 2021.

[Ca21] Carlini, N.; Tramèr, F.; Wallace, E.; Jagielski, M.; Herbert-Voss, A.; Lee, K.; Roberts, A.; Brown, T.B.; Song, D.X.; Erlingsson, Ú.; Oprea, A.; Raffel, C.: Extracting Training Data from Large Language Models. USENIX Security Symposium, 2021.

[Ch20] Chai, H.; Zhao, W.; Eger, S.; Strube, M.: Evaluation of Coreference Resolution Systems Under Adversarial Attacks. In: Proceedings of the First Workshop on Computational Approaches to Discourse. Association for Computational Linguistics, Online, S. 154-159, Nov. 2020, url: https://aclanthology.org/2020.codi-1.16.

[Dw06] Dwork, C.: Differential Privacy. In (Bugliesi, M.; Preneel, B.; Sassone, V.; Wegener, I., Hrsg.): Automata, Languages and Programming. Springer Berlin Heidelberg, Berlin, S. 1-12, 2006.

[Eb21] Ebers, M.; Hoch, V.; Rosenkranz, F.; Ruschemeier, H.; Steinrötter, B.: Der Entwurf für eine EU-KI-Verordnung: Richtige Richtung mit Optimierungsbedarf. Eine kritische Bewertung durch Mitglieder der Robotics & AI Law Society (RAILS), RDi 2/11, S.528-537, 2021.

[ES21] Ebert, A.; Spieker gen. Döhmann, I.: Der Kommissionsentwurf für eine KI-Verordnung der EU. Die EU als Trendsetter weltweiter KI-Regulierung, NVwZ 40/16, S. 1188-1193, 2021.

[GW22] Garat, D.; Wonsever, D.: Automatic Curation of Court Documents: Anonymizing Personal Data. Information 13/27, 2022, url: https://doi.org/10.3390/info13010027.

[Ha21] Habernal, I.: When differential privacy meets NLP: The devil is in the detail. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Online und Punta Cana, S. 1522-1528, Nov. 2021, url: https://aclanthology.org/2021.emnlp-main.114.

[Ha22] Habernal, I.: How reparametrization trick broke differentially-private text representation learning. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Dublin, (to appear), 2022, url: https://arxiv.org/abs/2202.12138.

[HPS17] Hagen, M.;Potthast, M.; Stein B.:Overview of the Author Obfuscation Task at PAN 2017: Safety Evaluation Revisited. In (Cappellato, L.; Ferro, N.; Goeuriot, L.; Mandl, T., Hrsg.): Working Notes of CLEF 2017 – Conference and Labs of the Evaluation Forum, Dublin, September 11-14, 2017. Bd. 1866. CEUR Workshop Proceedings, CEUR-WS.org, 2017.

[KB20] Kaulartz, M.; Braegelmann, T.: Rechtshandbuch Artificial Intelligence und Machine Learning. Beck C. H., 2020.

[KGD21] Krishna, S.; Gupta, R.; Dupuy, C.: ADePT: Auto-encoder based Differentially Private Text Transformation. In: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. Association for Computational Linguistics, Online, S. 2435-2439, Apr. 2021, url:

https://aclanthology.org/2021.eacl-main.207.

[LHZ18] Lee, K.; He, L.; Zettlemoyer, L.: Higher-Order Coreference Resolution with Coarse-to-Fine Inference. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers). Association for Computational Linguistics, S. 687-692, 2018.

[MSS20] Mahmood, A.; Shafiq, Z.; Srinivasan, P.: A Girl Has A Name: Detecting Authorship Obfuscation. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Online, S. 2235-2245, Juli 2020, url: https://aclanthology.org/2020.acl-main.203.

[NK20] Niemann, F.; Kevekordes, J.: Machine Learning und Datenschutz (Teil 1): Grundsätzliche datenschutzrechtliche Zulässigkeit. Computer und Recht 36/1, S. 17-25, 2020, url: https://doi.org/10.9785/cr-2020-360110.

[Pa18] Papernot, N.; Song, S.; Mironov, I.; Raghunathan, A.; Talwar, K.; Erlings-son, Ú.: Scalable Private Learning with PATE. In: 6th International Conference on Learning Representations, ICLR 2018, Vancouver, April 30 - May 3, 2018, Conference Track Proceedings, 2018.

[Pa20] Papastefanou, S.: „Database Reconstruction Theorem“ und die Verletzung der Privatsphäre (Differential Privacy). Computer und Recht 36/6, S. 379-386, 2020, url: https://doi.org/10.9785/cr-2020-360610.

[RG21] Roßnagel, A.; Geminn, C.: Vertrauen in Anonymisierung. Regulierung der Anonymisierung zur Förderung Künstlicher Intelligenz, Zeitschrift für Datenschutz - ZD 11/9, S. 487-490, 2021.

[Ro18] Roßnagel, A.: Pseudonymisierung personenbezogener Daten, ZD 9/6, S.241-288, 2018.

[RW21] Roos, P.; Weitz, C. A.: Hochrisiko-KI-Systeme im Kommissionsentwurf für eine KI-Verordnung. IT- und produktsicherheitsrechtliche Pflichten von Anbietern, Einführern, Händlern und Nutzern, MMR 24/11, S. 844-850, 2021.

[SA20] Schweter, S.; Akbik, A.: FLERT: Document-Level Features for Named Entity Recognition, 2020, arXiv: 2011.06993 [cs.CL].

[SB21] Spiecker gen. Döhmann, I.; Bretthauer, S., Hrsg.: Dokumentation zum Datenschutz, 84.Aufl. Baden-Baden 2021.

[Sc20] Schleipfer, S.: Pseudonymität in verschiedenen Ausprägungen. Wie gut ist die Unterstützung der DS-GVO? ZD 10/6, S. 284-290, 2020.

[SHB21] Schröder, F.; Hatzel, H. O.; Biemann, C.: Neural End-to-end Coreference Resolution for German in Different Domains. In Proceedings of the 17th Conference on Natural Language Processing (KONVENS 2021), S. 170-181, Düsseldorf, 2021.

[SHS19] Simitis, S.; Hornung, G.; Spiecker Döhmann, I., Hrsg.: Datenschutzrecht, DSGVO mit BDSG. Nomos, Baden-Baden, 2019.

[SIH21] Senge, M.; Igamberdiev, T.; Habernal, I.: One size does not fit all: Investigating strategies for differentially-private learning across NLP tasks. CoRRabs/2112.08159/, 2021, arXiv: 2112.08159, url: https://arxiv.org/abs/2112.08159. [St19] Struß, J. M.; Siegel, M.; Ruppenhofer, J.; Wiegand, M.; Klenner, M.: Overview of GermEval

Task 2, 2019 Shared Task on the Identification of Offensive Language. In: Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019). German Society for Computational Linguistics & Language Technology, Erlangen, S. 354-365, 2019.

[St20] Steinebach, M. et al.: Kapitel 4: Automatisierte Erkennung von Desinformationen. In: Steinebach, M.; Rinsdorf, L.; Krämer, N.; Roßnagel, A.: Desinformation aufdecken und bekämpfen: Interdisziplinäre Ansätze gegen Desinformationskampagnen und für Meinungspluralität. Baden-Baden: Nomos Verlagsgesellschaft (Schriften zum Medien- und Informationsrecht 45). S. 101-148, 2020.

[Sy18] Sydow, G.: Hrsg.: Europäische Datenschutzgrundverordnung, Handkommentar, 2. Aufl., Baden-Baden: Nomos Verlagsgesellschaft, 2018.

[WBH19] Winter, C.; Battis, V.; Halvani, O.: Herausforderungen für die Anonymisierung von Daten. In (David, K. et al., Hrsg.). INFORMATIK 2019. Gesellschaft für Informatik e.V., Bonn, S. 9-52, 2019.

[We20] Wu, W.; Wang, F.; Yuan, A.; Wu, F.; Li, J.: CorefQA: Coreference resolution asquery-based span prediction. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Online, S. 6953-6963, 2020.